Modeling linguistic evolution: a look under the hood

Cathcart, Chundra Aroor

Abstract: This paper takes a detailed look at some popular models of evolution used in contemporary diachronic linguistic research, focusing on the continuous-time Markov model, a particularly popular choice. I provide an exposition of the math underlying the CTM model, seldom discussed in linguistic papers. I show that in some work, a lack of explicit reference to the underlying computation creates some difficulty in interpreting results, particularly in the domain of ancestral state reconstruction. I conclude by adumbrating some ways in which linguists may be able to exploit these models to investigate a suite of factors which may influence diachronic linguistic change.

DOI: https://doi.org/10.1515/lingvan-2017-0043

Chundra Aroor Cathcart*

Modeling linguistic evolution: a look under the hood

https://doi.org/10.1515/lingvan-2017-0043
Received September 28, 2017; accepted December 6, 2017

Abstract: This paper takes a detailed look at some popular models of evolution used in contemporary diachronic linguistic research, focusing on the continuous-time Markov model, a particularly popular choice. I provide an exposition of the math underlying the CTM model, seldom discussed in linguistic papers. I show that in some work, a lack of explicit reference to the underlying computation creates some difficulty in interpreting results, particularly in the domain of ancestral state reconstruction. I conclude by adumbrating some ways in which linguists may be able to exploit these models to investigate a suite of factors which may influence diachronic linguistic change.

Keywords: evolutionary linguistics; diachronic linguistics.

1 Introduction

Computational biological methods designed to quantify the dynamics of evolution are rapidly gaining currency in diachronic linguistics, making highly visible and at-times controversial inroads into the subfield. Evolutionary methods are hardly the first quantitative methods to be employed in the subfield, but address certain aspects of diachronic change that cannot be treated in an explicit fashion by classical statistical methods. Typology, for instance, has long used quantitative methods to characterize the distribution of linguistic features across languages, often inferring information regarding diachronic processes from synchronic distributions. In general, this work involves controls for various sources of non-independence or bias in the cross-linguistic samples employed, such as shared genetic or geographic affiliation. Pioneering work in quantitative typology (e.g., Nichols 1986) makes use of frequentist statistical tests; while intuitive, classical statistical methods encounter problems when used to model data which violates the independence assumption. This shortcoming has generally been dealt with via principled exclusion of languages which may create bias in favor of a particular outcome (cf. Dryer 1989), though techniques of this sort are dissatisfying in that they require the available data to be underused, though mixed-effects models can be used to account for genetic skews in linguistic data.

The biological subfield known as PHYLOGENETIC COMPARATIVE METHODS contains a number of methodologies designed to quantify the distribution of traits across populations whilst accounting for non-independence between samples from related populations, an issue known as GALTON'S PROBLEM (Narroll 1961, Greenhill 2015). Some simpler methodologies treat genetic similarity as a nuisance factor which must be factored out in statistical tests; these methods do not, however, reveal much about (and make fairly neutral assumptions regarding) the evolutionary dynamics which may have led to contemporary differences across populations. For this reason, when investigating questions about the propensity of certain features across species or languages, biologists and linguists alike increasingly favor explicit phylogenetic approaches, particularly those which involve continuous-time Markov (CTM) processes of trait evolution. These methods usually involve the assumption that the evolution of a feature can be characterized by a stochastic process with parameters governing the distribution over time spans between events (e.g., gain, loss).

The use of these methods in linguistics is hardly uncontroversial. But regardless of their drawbacks, they have value in that they provide an explicit probabilistic framework for formulating and testing hypotheses

***Corresponding author: Chundra Aroor Cathcart,** Department of Linguistics and Phonetics, Lund University, Lund, Sweden; and Department of Comparative Linguistics, University of Zurich, Zurich, Switzerland, E-mail: chundra.cathcart@uzh.ch.
http://orcid.org/0000-0002-3066-4532

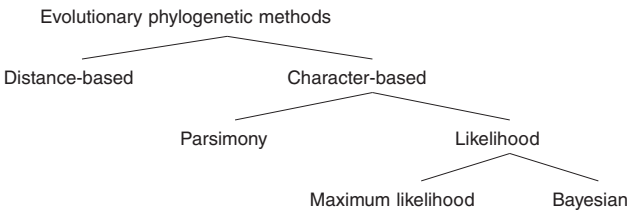

Figure 1: Schema of evolutionary phylogenetic methods.

regarding language change. It is certain that these methods will not be able to answer all questions asked by historical linguistics, particular since, in their current form, they abstract away from fine-grained details regarding usage, as well as individual- and community-level dynamics of language change. Still, a number of questions can be rather elegantly operationalized in terms of these methodologies.

This paper is not intended as a survey of all computational phylogenetic methods and their applications to linguistics, as a number of such surveys exist. My objective here is far more humble: I attempt to provide an explicit look at the math which underlies the computation of tree likelihoods, and how this information can be used to make inferences about evolutionary dynamics of features over a tree, as well as an overview of some work applying these methods to linguistic phenomena. I outline a few pitfalls encountered in assessing the performance of models of this sort, and hope to drive home the point that linguists cannot rely solely on computational biological software packages to address all interesting questions in quantitative diachronic linguistics, and must not allow these methodologies to serve as a black box. It is hoped that ongoing collaborative software development among linguists will serve this end.

2 Computational phylogenetics: an overview

There are already several comprehensive and thorough tutorials (tailored to linguists) addressing the inference of tree topologies (Nichols and Warnow 2008; Dunn 2015), the latter of which touches on comparative phylogenetic methods as well. This review makes no attempt to supersede these works, but it is necessary to rehash some of the terminology and distinctions found in these surveys. Computational cladistic methods used in evolutionary biology can be divided into the taxonomy seen in Figure 1. The highest-order division is generally one drawn between distance-based and character-based methods. The former method takes as its starting point a matrix of pairwise distances between taxa (i.e., species or languages) calculated on the basis of the features they exhibit, and clusters the taxa into smaller clades. These pairwise distances represent the degree of difference between two taxa according to a set of feature or trait values.

Character-based methods return a set or sample of trees representing possible evolutionary histories of a set of characters or traits over the tree. Of these, a tree is usually chosen (or constructed from the sample) which fulfills some objective, either probabilistic (i.e, it exists in a region of high likelihood) or otherwise (e.g., parsimony methods, which favor trees minimizing the total number of character-state changes). Likelihood methods allow for a larger range of probabilistic models of character evolution, far beyond models that simply aim to minimize the number of changes or parallel changes over the tree.[1] As in other domains of statistics, maximum likelihood estimation seeks values for unknown quantities or parameters which maximize the likelihood (i.e., the probability of the data given the parameters), while Bayesian methods assume a prior distribution over parameters, and approximate a posterior probability distribution for the parameter in question (i.e., a distribution characterizing the probability mass or density of the parameter given the data) via a Markov chain Monte Carlo (MCMC) procedure.

1 Attempts have been made to formulate parsimony approaches in terms of likelihood; for a discussion of the statistical properties of parsimony, see Felsenstein (2004, Ch. 9). This paper focuses chiefly on likelihood methods.

A distinctive feature of character-based methods (which specify an explicit model regarding the evolutionary behavior of the traits in question) is that given an extant tree topology, trait histories can be inferred; characters can be mapped independently onto a tree or set of trees that have been inferred via the sorts of procedures described in Nichols and Warnow (2008) and Dunn (2015). This allows us to (among other things) reconstruct the probability of a feature state at a given internal node.

3 Continuous-time Markov models of discrete character evolution

Likelihood methods make it possible to compute the probability that observed data found among a set of taxa (i.e., languages) were generated by an evolutionary model over a tree of some topology. Under a common type of evolutionary model, discrete traits are thought to evolve according to a continuous time Markov (CTM) process, transitioning from one state to another with a certain probability over a given interval of time. These models range from very simple to highly complex, in terms of the free parameters assumed. A basic, single-parameter model of binary trait evolution assumes that a trait evolves according to one symmetric mutation rate that governs the change $0 \rightarrow 1$ as well as $1 \rightarrow 0$. A slightly more complex model assumes that a binary character evolves according to two different parameters, a gain rate (governing the change $0 \rightarrow 1$) and a loss rate (governing the change $1 \rightarrow 0$). If we know the gain and loss rates, it is trivial to calculate the probability that the trait in question is present at an interior node of the tree under a CTM process. These probabilities are usually computed via matrix exponentiation, which can be costly for large numbers of rates (i.e., transitions between multistate, i.e., non-binary, values of a trait). For two rates (e.g., a gain rate α and a loss rate β), we can easily compute the continuous-time Markov chain transition matrix (for discussion and derivation, see Liggett 2010:59–60), with each cell representing $p_{ij}(t; \alpha, \beta)$, the probability of transitioning from state i (the ancestral state) to state j (the descendant state) over a branch of length t, given the rates α, β:

(1)

	$j = 0$	$j = 1$
$i = 0$	$\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t}$	$\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t}$
$i = 1$	$\frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)t}$	$\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-(\alpha+\beta)t}$

In the biological literature, authors rarely provide the actual matrix of transition probabilities, but give the instantaneous rate matrix from which it is derived, which for two rates generally looks like one of the following (the exact convention may vary from author to author):

$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \qquad Q = \begin{pmatrix} - & \alpha \\ \beta & - \end{pmatrix}$$

For continuous-time Markov processes with more than two rates, transition probability matrices like that seen in (1) must be computed via matrix exponentiation; alternative approaches for estimating these matrices have been suggested as well (on the use of uniformization, one such approach, in biology, see Irvahn and Minin 2014).

3.1 Computing the likelihood of a tree under an evolutionary model: the pruning algorithm

At the outset, we do not know the latent parameters α and β (even if we can assume a fixed tree topology), and must find values which maximize the likelihood of the tree (under a Maximum Likelihood approach), or

values in regions of high posterior probability (under a Bayesian approach). This process involves computing the probability of the tree, which requires summing over all possible state combinations of the interior nodes of the tree; this number increases exponentially with the number of interior nodes in the tree, making it computationally intractable for large trees. However, the independence structure of the tree — the fact that the state of a node depends only on the state of its direct ancestor and the length of the branch along which it evolves — can be exploited to efficiently compute this likelihood according to the PRUNING ALGORITHM (Felsenstein 2004:253–5), which starts at the tips of the tree (i.e., the terminal nodes, which represent languages with observed data) and recursively computes the likelihood that each interior node of the tree is in state 0 or 1 (in the case of a binary character), given a tree topology and some values for α and β (which I write $\ell_{\alpha,\beta}(\mathbf{n} = i)$; this notation varies widely across the literature on biology and graphical statistics), given the likelihood that each of its children are in either state 0 or 1 ($\mathbf{branch}_{n,\mathbf{child}}$ denotes the length of the branch connecting n to one of its children).

$$\ell_{\alpha,\beta}(\mathbf{n} = i) = \prod_{\mathbf{child} \in \mathbf{children\ of\ n}} \left(\sum_{j=0}^{1} p_{ij}(\mathbf{branch}_{n,\mathbf{child}}; \alpha, \beta) \ell(\mathbf{child} = j) \right)$$

At the tips of the tree, $\ell(\mathbf{state(n)} = i)$ is equal to either 0 or 1 (provided that data are not missing). The overall likelihood of the tree is computed via the sum of the probabilities that the root is in each state (given its direct descendants) weighted by the stationary probability of each state[2]:

$$\ell_{\alpha,\beta}(\mathbf{tree}) = \frac{\alpha}{\alpha+\beta} \ell_{\alpha,\beta}(\mathbf{root} = 1) + \frac{\beta}{\alpha+\beta} \ell_{\alpha,\beta}(\mathbf{root} = 0)$$

The pruning algorithm is also a crucial ingredient in searching the space of possible tree topologies for regions of higher likelihood.

Yang (2014:133–143) gives an extensive survey of maximum likelihood methods for finding rates of evolution. In the case of a single, symmetric rate of evolution, the maximum likelihood estimate can be found via a simple line search, since only one variable is involved. For more than one rate, the math is slightly more complicated. MCMC approaches do not require much in the way of complicated math; different values for the rates in question are iteratively proposed, and accepted stochastically if they improve the likelihood of the tree (this procedure is described in greater detail below). MCMC approaches are generally Bayesian, and return posterior distributions over rates.

Via Bayes' Rule, the relationship between the posterior probability $p(\theta|D)$, the likelihood $p(D|\theta)$, the prior $p(\theta)$, and the marginal likelihood is the following (θ denotes the unknown parameter[s], while D denotes the observed data):[3]

$$p(\theta|D) = \frac{\text{LIKELIHOOD} \times \text{PRIOR}}{\text{MARGINAL LIKELIHOOD}} = \frac{p(D|\theta)p(\theta)}{p(D)}$$

The marginal likelihood can be difficult to compute, as it requires summing over or integrating out the entire hypothesis space ($p(D) = \sum_\theta p(D, \theta) = \sum_\theta p(D|\theta)p(\theta)$, if θ is discrete); since it does not vary from hypothesis to hypothesis, it can be ignored, as the relative probability of a hypothesis is as important (for most practical purposes) as its absolute probability.

While inferring distributions for α and β via MCMC, we care about the relative posterior probabilities (which, in this case, are equivalent to the relative likelihoods) of different values of α and β. It is standard

2 For the binary feature described here, the stationary probabilities of presence and absence, respectively, work out to $\frac{\alpha}{\alpha+\beta}$, $\frac{\beta}{\alpha+\beta}$; technically speaking, the stationary probabilities of states in a continuous-time Markov chain are comprised by a vector $\boldsymbol{\pi}$ which satisfies the equations $\boldsymbol{\pi}Q = 0$ and $\sum \boldsymbol{\pi} = 1$.

3 The notation $p(y|x)$ denotes the probability of y given x, and is equivalent to $\frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\sum_y p(x,y)}$ (if y is discrete) or $\frac{p(x,y)}{\int p(x,y)dy}$ (if y is continuous).

to randomly initialize α and β; the likelihood of the tree under the current values of α and β ($\ell_{\alpha,\beta}(\mathbf{tree})$) is computed via the pruning algorithm, and new values $\alpha^{'}$, $\beta^{'}$ are proposed, drawn from Gaussian distributions centered around α, β:

$$\alpha^{'} \sim N(\alpha, \sigma_\alpha), \beta^{'} \sim N(\beta, \sigma_\beta)$$

The likelihood under $\alpha^{'}$, $\beta^{'}$, $\ell_{\alpha^{'},\beta^{'}}(\mathbf{tree})$, is computed as well. The proposed values $\alpha^{'}$, $\beta^{'}$ are accepted as new current values if a randomly generated number a (drawn from the Uniform distribution between 0 and 1) is less than $\frac{\ell_{\alpha^{'},\beta^{'}}(\mathbf{tree})}{\ell_{\alpha,\beta}(\mathbf{tree})}$.[4] Above, σ_α, σ_β denote the step sizes for the two parameters. These values are nontrivial, as they have an effect on the extent to which the space of possible states is explored; in general the state space is adequately traversed by a Markov chain when roughly 23.4% of proposed states are accepted (Rosenthal 2011). After a period of time, the algorithm can be expected to draw samples from the posterior distributions of α and β.

Inference is generally run on three or four chains, all of which are initialized separately at different values, but under optimal circumstances are ultimately expected to "mix," i.e., to sample from the same posterior distribution. An initial quantity of samples (in conservative practice, the first half) is generally discarded (the so-called "burn-in"). The chains' convergence can be monitored using the Potential Scale Reduction Factor $\hat{R}$ for each posterior sample, which is the square root of the ratio of the between-chain variance and the average within-chain variance, under the assumption that the posterior distribution is normal (Gelman and Rubin 1992). Values below 1.1 indicate good convergence.

3.2 Ancestral state reconstruction and stochastic character mapping

Once maximum likelihood estimates or posterior distributions for α and β are obtained, it becomes possible to reconstruct the probability that a feature is present at internal nodes of the tree, and also to simulate mutations along branches, a process known as stochastic character mapping (Nielsen 2002; Bollback 2006). First, the probability that the feature is present at the root is calculated:

$$p(\mathbf{root} = 1|\alpha, \beta) = \frac{\frac{\alpha}{\alpha+\beta}\ell_{\alpha,\beta}(\mathbf{root} = 1)}{\frac{\alpha}{\alpha+\beta}\ell_{\alpha,\beta}(\mathbf{root} = 1) + \frac{\beta}{\alpha+\beta}\ell_{\alpha,\beta}(\mathbf{root} = 0)}$$

A Bernoulli variate parameterized by this probability is drawn, yielding a value $i = \{1, 0\}$. Subsequently, for each child n of the root, the probability that the feature is present at n given the sample value i for the root can be computed:

$$p(\mathbf{n} = 1|\mathbf{root} = i, \alpha, \beta) = \frac{\ell_{\alpha,\beta}(\mathbf{n} = 1)p_{i1}(\mathbf{branch}_{\mathbf{root},n}; \alpha, \beta)}{\ell_{\alpha,\beta}(\mathbf{n} = 1)p_{i1}(\mathbf{branch}_{\mathbf{root},n}; \alpha, \beta) + \ell_{\alpha,\beta}(\mathbf{n} = 0)p_{i0}(\mathbf{branch}_{\mathbf{root},n}; \alpha, \beta)}$$

Again, a Bernoulli variate is drawn; this procedure carries on down the tree until the nodes directly above the tips are reached. After state values have been simulated at the interior nodes of the tree, the birth and death locations can be simulated along branches by drawing values from exponential distributions parameterized by α and β, respectively.

4 This update is known the Metropolis-Hastings algorithm (in this case equivalent to the Metropolis algorithm, since the proposal distribution is Gaussian and therefore symmetric, e.g., $p(\alpha^{'}|\alpha, \sigma_\alpha) = p(\alpha|\alpha^{'}, \sigma_\alpha)$)

3.3 Extensions and linguistic applications

The CTM model of character evolution and the pruning algorithm serve as the backbone for a large number of models of evolution which have enjoyed use in linguistics. The model described above is among the most simple; it assumes a "strict clock," i.e., that across all branches in the tree, α and β are drawn from the same global distributions. This simplicity and restrictiveness make for tractable and fast inference and prevent against overfitting, but the assumptions involved may not be entirely realistic. A "relaxed clock" assumes that each branch has its own individual rate parameter, which can be multiplied by the global change rates.[5] The covarion family of models assumes that there are regions of slow and fast change over the tree; in some formulations, the slow rates are usually equivalent to the non-slow gain and loss rate multiplied by some constant between 0 and 1 (Wang et al. 2006). Hidden rates approaches extend this model to more than two rate classes (Beaulieu and O'Meara 2014).

Methodologies which infer evolutionary rates have enjoyed some use to date in diachronic linguistics. Dediu (2010) uses (among other methods) a two-rate model to estimate the stability of typological features in a number of families, quantifying stability according to the gain rate relative to the loss rate. Chang (2014) introduces a latent-state character model suited to the evolution of root-meaning characters. Under this model, a character is born at a certain rate, after which point it remains "latent" for a period of time; it becomes "active" at a certain rate, or dies; if active, it dies at a certain rate.

Pagel (1994) and Pagel and Meade (2006) present methods designed to measure co-evolutionary dependencies between a pair of traits. Two inference procedures are carried out: first, an independent model is run, where each trait is gained or lost independently according to a CTM process with four rates (two per trait). This is followed by a dependent model where the joint evolution of the two traits is modeled via a CTM process transitioning between four states representing all 2×2 combinations of presence and absence; the dependent model has eight rates, which treat all possible transitions between combinations differently (with the exception of the changes $0, 0 \rightarrow 1, 1$ and $1, 1 \rightarrow 0, 0$, the rates of which are set to zero). In Bayesian approaches to this procedure, evidence in favor of the dependent hypothesis versus the independent hypothesis is assessed according to the Bayes Factor of the two models (the ratio of their marginal likelihoods). This method provides an explicit approach to testing hypotheses regarding evolutionary associations between two features, provided that all possible combinations are observed at the tips of the tree. At the same time, Maddison and FitzJohn (2015) caution against making inferences regarding the causal relationship between the features under study; it could be the case that they are associated with each other because they are caused by a third factor (cf. Pearl 2009); furthermore, the authors note that this methodology will infer a correlation between features even if one of the features evolves only once. This technique is applied by Dunn et al. (2011) to the question of whether pairs of features involved in Greenbergian word order harmonies exhibit dependent evolution, with a negative result.

We may additionally desire to uncover correlated evolution between more than two traits. However, it is impossible to run the procedure described above for all possible combinations of a large number of traits. Reversible-jump MCMC provides a solution to this problem, allowing parameters pertaining to pairwise correlation between traits to be added or deleted over the course of the inference procedure. Haynie and Bowern (2016) utilize this method to analyze trajectories in the gain and loss of color terms over a posterior tree sample taken from Bowern and Atkinson 2012, finding broad support for Berlin and Kay (1969) theory that color terms are gained in a partially fixed order, but find also evidence for alternative pathways (e.g., green gained in the absence of red) and loss of color terms.

At the time of the publication of Dunn 2015, relatively little linguistic research had made use of ancestral state reconstruction and stochastic character mapping, a fact noted in the author, with some exceptions such as Maurits and Griffiths 2014, a cross-family study of the evolution of word order, in which the authors discuss

5 It is worth noting that if one models character evolution over a tree sample rather than a fixed tree, the branch lengths, which vary across the sample, take on the task of introducing stochasticity that would otherwise fall upon the relaxed clock model. One can also simulate this stochasticity using a fixed tree by adding some random noise to the branch lengths.

how choice of prior affects ancestral state reconstruction at the root, among other issues. The state of affairs has changed since then.

A handful of papers have investigated character histories over tree topologies using parsimony. Jäger and List (2016) use parsimony to reconstruct root-meaning traits to Proto-Indo-European and Proto-Austronesian, and evaluate the performance of their model explicitly using metrics that are commonplace in other branches of computational linguistics. The authors adopt a strategy of comparing computationally reconstructed root-meaning traits against the "list of cognate classes present in the proto-language according to expert assessments" (p. 3). From this wording, it is not clear to me whether an automatically reconstructed root-meaning trait is evaluated against whether the expert considers the root to exist in the proto-language with a particular meaning, or whether the expert simply considers the root to exist in the proto-language. If the former, then all caveats regarding the difficulty of semantic reconstruction apply. If the latter, a root-meaning reconstruction could potentially be judged correct simply because the expert believes the root to be present in the proto-language, but with some other meaning. This could potentially lead to inflated recall values. Given the fact that the recall values reported by the authors never exceed 0.734 (still a relatively high number), it seems unlikely that the latter potentially anti-conservative metric was employed.

It is useful to have a baseline established via expert assessments against which to measure reconstruction methodologies. However, experts often disagree in their views; for this reason, it is as crucial to know how a qualitative conclusion was reached as it is how a quantitative result was achieved. It may be the case that a philologist carries out reconstruction effectively implements the parsimony algorithm by hand, i.e., by applying Occam's Razor and attempting to avoid recurrent changes (though the scores that the authors report show that these processes are not 100% identical), such that this accuracy might not be particularly surprising. It is worth noting that the authors seem to treat this exercise as more of a sanity check — a step along the way to future work detecting large-scale recurrent change between branches.

Additionally, some research has carried out ancestral state reconstruction using likelihood-based methodologies. Recent work by Widmer et al. (2017) uses stochastic character mapping over a tree sample of Indo-European languages to demonstrate that at least one strategy of NP recursion was available over the course of Indo-European prehistory, according to their evolutionary model; ancestral state reconstruction is a crucial ingredient in this analysis.

Dunn et al. (2017) investigate the evolution of morphological case-marking on the subjects of a large number of verbs in Germanic languages; verbs in many Germanic languages co-occur with morphologically dative and accusative as well as nominative subjects. The study, based on a large amount of carefully curated data, is the first in linguistics to make use of (Pagel 1999) MULTISTATE methodology. Each verb is treated as a non-binary character with values denoting how the subject is marked, i.e., with N(ominative), A(ccusative), or D(ative) case. The authors use a handcrafted tree topology which they claim represents the consensus of historical linguists; however, a subgroup like Germanic, where internal language contact is well-documented, is an area where a tree sample, rather than a fixed tree, would be highly welcome, since there is arguably a degree of phylogenetic uncertainty. The authors compare an unconstrained model, where transitions between all three strategies occur at non-zero rates, against a set of models which set certain transition rates to zero in order to constrain the evolutionary pathway between states. They find that while a model representing "dative sickness" (where transitions A $\rightarrow$ D, A $\rightarrow$ N, and D $\rightarrow$ N are permitted, but not transitions D $\rightarrow$ A, N $\rightarrow$ D, or N $\rightarrow$ A) is not favored in terms of likelihood, it reconstructs ancestral state values which agree completely with qualitative judgments made by the authors and other specialists; they interpret this apparent success as evidence in favor of the dative sickness model.

If one wishes to see whether a particular model makes reconstructions which agree with a human expert (in order to confirm that such a model is appropriate to assume for orthogonal work, as do Jäger and List), that is indeed reasonable. However, the degree to which ancestral state reconstruction agrees with a philological reconstruction should have no bearing on model selection. This accuracy no way serves as independent evidence in favor of the dative sickness model, as Dunn et al. (2017) seem to imply; the authors have built their a priori assumptions regarding directionality of change in case assignment into the CTM model they use. When we look at reconstructions made by the unconstrained model (p. e15), we see that N subjects are

erroneously reconstructed with high probability for a number of verbs, but for the dative sickness model, N subjects are reconstructed with zero probability; this appears to account for a large part of the model's accuracy, as assessed by the authors. These facts should be unsurprising: in the dative sickness model, the rates for N → A and N → D change are set to zero. This means that N can be reconstructed with nonzero likelihood at an interior node of the tree only if N is present at all of the tips that descend from the node in question. In the dative sickness model, the D → A transition rate is also set to zero; hence, if D, A, and N are attested at the tips of the tree, A will be reconstructed at the root with 100% probability. If only D and A are present, the same is true. If only D and N are present, D is most likely to be reconstructed, but if the rates A → D and A → N are positive, there is still a small chance of A being reconstructed. And indeed, D is reconstructed only for verbs where there are no unambiguous instances of A at the tips of the tree; otherwise, A is reconstructed with 100% probability. It is pleasing when the result of a quantitative methodology dovetails with a philological analysis — evaluating the results of ancestral state reconstruction according to expert assessments provides a necessary sanity check for linguists — but in many situations, it behooves us to investigate exactly why this is the case. A simpler example is as follows: if, for instance, we wished to use a similar model to reconstruct word order for a family of languages exhibiting SOV and SVO word order, but chose to set the rate of SVO → SOV change to zero, we could expect SOV word order to be reconstructed with 100% probability. It may be that we believe SOV to be the ancestral word order, but this is not an observable fact against which to evaluate whether the model we have chosen is "correct."

This discussion raises, it is hoped, some important questions regarding what it is exactly that we wish to *do* with CTM methods, ancestral state reconstruction, and related methodologies. It is important to recall that there is more than one way to allocate credibility to a particular hypothesis. The authors' dative sickness model, which set certain transition rates to zero, is not favored in terms of likelihood. But it may be the case that transitions of the type D → A, N → A, and N → D occur with a probability that is greater than zero, but considerably lower than the transitions A → D, A → N, D → N. After all, the authors do provide some examples of dative to accusative change (p. e6), and it is therefore puzzling why they favor a model which fixes the probability of this type of change at zero. It could be the case that evidence can be adduced in favor of a weak formulation of the authors' dative sickness model, one which predicts that D → A, N → A, and N → D changes happen at a less frequent rate than other changes. In theory, this could be done by running the unconstrained model for all of the verbs in the dataset, and then inspecting the rate estimates to see whether this pattern holds (the authors do not make the resulting MLEs available).

I attempted to investigate this question using the data and tree from the paper, which the authors have made publicly available. I used MCMC to sample from the log posterior distributions of all six rates, constraining parameter space such that evolutionary rates did not exceed ≈ 2.7 mutations per 1000 years. The code used for inference can be found at https://github.com/chundrac/vanguard. In general, I found that this procedure did not always lead to convergence, and that chain acceptance rates often exceeded their recommended values (see above); this was undoubtedly an artifact of carrying out MCMC for a large number of rates over a relatively small tree, with a somewhat constrained state space. For this reason, we cannot necessarily expect to obtain the same rate values across different random number seeds. Median values for each rate, aggregated across verbs, are shown in the boxplot in Figure 2. The mean of the median log values of the "favored" rates (qAD, qDN, qAN) is slightly higher than that of the "disfavored" rates (−23.8 vs. −28.1), but this difference is not significant (Mann-Whitney $U = 698, p = .58$). This is undoubtedly due to the fact that values for the rate qDA, included among the disfavored rates, are relatively high; while lower than that of qAD (−30.5 vs. −24.8), it is not significantly so. The rate qAN is also higher than qNA (−21.3 vs. −25.5), but this difference is insignificant as well. The same is true for qDN as opposed to qND (−25.2 vs. −28.2); like the others, this difference is insignificant. All in all, the rates for trajectories favored by the sickness model are higher than their counterparts in the opposite direction, but there is no significant evidence in favor of these asymmetries. Given the fact that other quantitative studies have found relatively strong support for the dative sickness hypothesis (del Prado Martín and Brendel 2016), evolutionary methods may not be the approach best suited to this question.

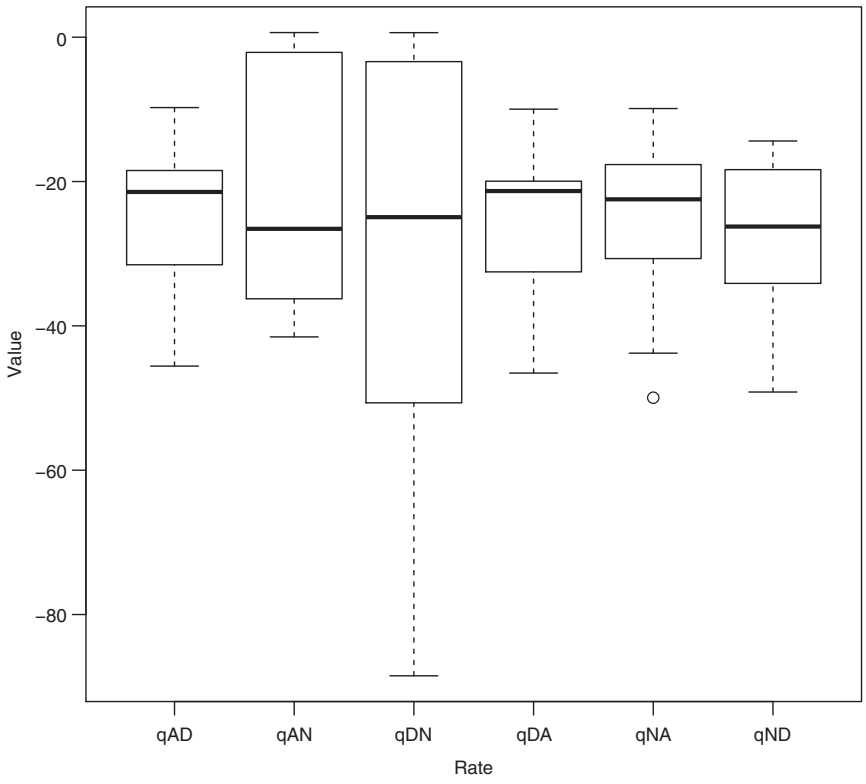

Figure 2: Median values for log rates, aggregated across verbs.

4 Conclusion

The majority of methodological papers cited come with their own free software. At the time of writing, many widely used computational biological software packages have interfaces with programming languages like R. A basic R package for phylogenetic analysis is **ape** (Paradis et al. 2004); many of this package's conventions, including the **phylo** format for representing trees, serve as the backbone for packages that have been subsequently developed, such as **picante** (Kembel et al. 2010), which contains a wide array of comparative methods. The **phytools** package (Revell 2012) also incorporates a number of commonly used comparative methods, including several of those discussed above; this package has excellent documentation[6] and is frequently updated.

At the same time, the software tools mentioned above are designed for biological research, and have built-in conventions which may be inconvenient for linguists, although, certain computational phylogenetic models have been designed with linguistic uses in mind (Nicholls and Gray 2006); it is hoped that as linguistic use of comparative phylogenetic tools increases, linguists can play a greater role in shaping the development of these methodologies, and at the time of writing, the number of computational phylogenetic tools geared primarily toward linguists is increasing in real time (e.g., List and Forkel 2016; Maurits 2016). In general, researchers in data-driven historical linguistics have been diligent in making their data and code available and transparent on repositories such as GitHub. The developers of RevBayes, a program for phylogenetic graphical models (Höhna et al. 2016), have transparency and flexibility in mind.

As mentioned above, many linguists tend to be resistant to the use of phylogenetic models because they take issue with several of the simplifying assumptions that these models make. It will increasingly become the

6 http://www.phytools.org/

duty of interested linguists to tweak a number of the assumptions in question. In some cases, these issues can be addressed within the confines of off-the-shelf biological software packages. Others will require the production of new software or code tailored to linguistic questions. This process will undoubtedly be a messy and difficult one but will ultimately enrich the field.

Acknowledgment: I am grateful to Claire Bowern, Gerd Carling, Erich Round and Michael Weiss for helpful comments on and discussion of various versions of this paper. Any errors or omissions are my own.

References

Beaulieu, Jeremy M. & Brian C. O'Meara. 2014. Hidden Markov models for studying the evolution of binary morphological characters. In László Zsolt Garamszegi (ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice*, 395–408. Heidelberg, New York, Dordrecht, London: Springer.

Berlin, Brent & Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Berkeley, CA: University of California Press.

Bollback, Jonathan P. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7. 88.

Bowern, Claire & Quentin D. Atkinson. 2012. Computational phylogenetics and the internal structure of Pama–Nyungan. *Language* 88(4). 817–845.

Chang, William. 2014. A vanishing, multiple-gain lexical trait model: Challenges and opportunities in lexical data and analysis. Paper presented at the Workshop Towards a Global Language Phylogeny, Jena, 17–20 September. Available at `http://lingsoup.com/talk/jena-2014.pdf` (accessed 1 October 2015).

Dediu, Dan. 2010. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society of London B* 278(1704). 474–479.

del Prado Martín, Fermín Moscoso & Christian Brendel. 2016. Case and cause in Icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, 2421–2430. Association for Computational Linguistics.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292.

Dunn, Michael. 2015. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 190–211. New York & Oxford: Routledge.

Dunn, Michael, Tonya Kim Dewey, Carlee Arnett, Thórhallur Eythórsson & Jóhanna Barðdal. 2017. Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic. *Language* 93(1). e1–e22.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79–82.

Felsenstein, Joseph. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Gelman, Andrew & Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7. 457–511.

Greenhill, Simon J. 2015. Demographic correlates of language diversity. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 557–578. New York & Oxford: Routledge.

Haynie, Hannah J. & Claire Bowern. 2016. A phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences* 113(48). 13666–13671.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck & Fredrik Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65(4). 726–736.

Irvahn, Jan & Vladimir N. Minin. 2014. Phylogenetic stochastic mapping without matrix exponentiation. *Journal of Computational Biology* 21(9). 676–690.

Jäger, Gerhard & Johann-Mattis List. 2016. Investigating the potential of ancestral state reconstruction algorithms in historical linguistics. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*. Tübingen: University of Tübingen, online publication system, `https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68641` (accessed 1 April 2017).

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg & C. O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26. 1463–1464.

Liggett, Thomas M. 2010. *Continuous time Markov processes: an introduction*, vol. 113 Graduate Studies in Mathematics. Providence, RI: American Mathematical Society.

List, Johann-Mattis & Robert Forkel. 2016. Lingpy. A Python library for historical linguistics. `http://lingpy.org`. doi:https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

Maddison, Wayne P. & Richard G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64(1). 127–136.
Maurits, Luke. 2016. Beastling: a linguistics-focussed command line tool for generating beast xml files. Python package. https://github.com/lmaurits/beastling.
Maurits, Luke & Thomas Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences* 111(37). 13576–13581.
Narroll, Raoul. 1961. Two solutions to Galton's Problem. *Philosophy of Science* 28. 15–29.
Nicholls, Geoff K. & Russell D. Gray. 2006. Quantifying uncertainty in a stochastic Dollo model of vocabulary evolution. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 161–71. Cambridge: McDonald Institute for Archaeological Research.
Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62. 56–119.
Nichols, Johanna & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820.
Nielsen, Rasmus. 2002. Mapping mutations on phylogenies. *Systematic Biology* 51(5). 729–739.
Pagel, Mark. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B* 255. 37–45.
Pagel, Mark. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* 48(3). 612–622.
Pagel, Mark & Andrew Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by Reversible-Jump Markov Chain Monte Carlo. *The American Naturalist* 167(6). 808–825.
Paradis, E., J. Claude & K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20. 289–290.
Pearl, Judea. 2009. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Revell, Liam J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3. 217–223.
Rosenthal, Jeffrey S. 2011. Optimal proposal distributions and adaptive MCMC. In Steve Brooks, Andrew Gelman, Galin L. Jones & Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, 93–112. Boca Raton, FL: Chapman & Hall/CRC.
Wang, Huai-Chun, Matthew Spencer, Edward Susko & Andrew J. Roger. 2006. Testing for covarion-like evolution in protein sequences. *Molecular Biology and Evolution* 24(1). 294–305.
Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: evidence from Indo-European. *Language* 93(4). 799–826.
Yang, Ziheng. 2014. *Molecular evolution: A statistical approach*. Oxford: Oxford University Press.